Пименов Е. Н.

Англо-русский тезаурус по сохранности документов и некоторые особенности его составления

В технологии документальных баз данных и, в том числе, в работах по индексированию и составлению тезаурусов есть много различий, обусловленных тем, что имеется несколько типов БД и тем, что системы работают и создаются в различных условиях. Особенно много различий имеется между технологиями крупных универсальных автоматизированных библиотечно-информационных систем (АБИС) и, с другой стороны, небольшими и обычно специализированными документальными базами данных. Далее эти два класса систем мы будем именовать сокращениями МБД – от *малые базы данных* или *базы данных малого объема* и ББД – *большие БД* или *базы данных большого объема*. В этой работе рассматривается взаимодействие разных параметров и характеристик системы при разработке тезауруса по сохранности документов. В состав лингвистических средств базы данных по обеспечению сохранности документов (БД ОСД, содержит по состоянию на май 2001 г. около 12 тыс. библиографических записей) входят:

- семантически сильный информационно-поисковый тезаурус (ИПТ), включающий в настоящее время 5120 английских и русских ключевых слов (КС), из которых 1674 КС составляют дескрипторы;
- схема предметной области, имеющая вид тематического рубрикатора информации;
- детальная и хорошо апробированная методика индексирования документов;
- программа автоматизированного индексирования информации (АИ), разработанная в БАН для программной среды ППП CDS/ISIS.

Два указанных класса систем в основном различаются по таким их технологическим характеристикам.

Преимущественное назначение или использование документальной системы. За исключением БД типа АРМ библиографа, предназначенных для компьютерной подготовки библиографических указателей и распечатки библиографических карточек, основным назначением МБД является проведение *тематических поисков.* Проведение *адресных поисков* по заглавиям, авторам и другим форматным полям документов является более характерным для ББД и в малых системах является редкостью.

Информационно-поисковый язык базы данных. Основным языком индексирования информации в большинстве МБД являются жестко нормализованные или слабо нормированные ключевые слова. В крупных универсальных системах, и особенно – класса АБИС, для обработки и поисков информации в основном применяются ББК, УДК, рубрикатор ГАСНТИ и другие классификационные ИПЯ, а ключевые слова здесь обычно являются лишь вспомогательным средством информационного поиска. Указанные выше различия в преимущественном назначении баз данных влияют на глубину индексирования и степень нормализации КС. В МБД глубина индексирования и уровень нормализации лексики в ПОДах обычно являются более высокими, чем в крупных универсальных системах.

По нашему мнению, имеется тесная связь между основным ИПЯ базы данных и характером наиболее массовых информационных запросов в системе. Так, например, в МБД по огнеупорным изделиям и материалам [1], бывшей в промышленной эксплуатации около 18 лет, из 5 тыс. проведенных за данный период ретроспективных поисков информации, адресных поисков по фамилиям авторов и наименованиям первоисточников было не более 20, поиски по заглавиям не проводились, и практически не было поисков (кроме экспериментальных) по индексам УДК и МКИ. Из приведенной статистики вытекает, что большое количество адресных поисков, выполняемых в крупных АБИС, в действительности имеет, видимо, тематический характер. Так, адресный поиск по авторам или заглавиям работ может на самом деле быть тематическим, если он выполняется с целью подбора литературы по нужной читателю теме.

Объем обрабатываемой информации. При больших объемах работы, особенно если она выполняется в короткие сроки и при некомплекте штатов БД (а это – обычная ситуация в крупных системах), индексирование ключевыми словами, для сокращения его трудоемкости, может в значительной степени упрощаться и результаты индексационных работ могут быть недостаточно качественными. Малый объем обрабатываемой информации обусловливает многие преимущества, в частности, более высокое качество индексирования в специализированных МБД.

Состав документов. В большей части малых систем главная часть индексируемой информации (в некоторых случаях доходящая до 100 %) приходится на аналитические записи, а количество обрабатываемых книг обычно является не очень большим. Уровень специфичности содержания и индексирования текстов заглавий статей выше, чем уровень специфичности заглавий книжных изданий. Последним обычно

определяется большая глубина индексирования информации в МБД.

Особенности комплектования системы. В ББД, особенно в базах данных класса АБИС информация обрабатывается и, в том числе, индексируется с использованием первоисточников, то есть de vizu. Индексаторы в МБД очень часто работают со вторичными источниками информации в виде библиографических указателей, списков и картотек очень разного, иногда невысокого качества и основанных на разных стандартах, отечественных и зарубежных. Поскольку в этих условиях в малых системах может отсутствовать простая физическая возможность индексирования информации de vizu, здесь в более преимущественном положении находятся ББД.

Трудоемкость и стоимость документального ввода. По литературным данным средняя стоимость каталогизации одной книги (вместе с работой по индексированию информации) в Библиотеке Конгресса США оценивается в $50, а общие затраты на каталогизацию книг сравнимы с затратами на комплектование [2]. Затраты на ввод информации в МБД должны быть заведомо меньшими, а сокращение трудоемкости индексирования ключевыми словами в ББД иногда достигается путем уменьшения глубины или каких-то других показателей качества индексирования.

Штаты и квалификация персонала. Крупные базы данных имеют или по крайней мере должны иметь в своих штатах необходимое количество специалистов высокой квалификации: программистов, операторов подготовки данных, каталогизаторов, систематизаторов, специалистов по индексированию и разработке тезаурусов. В МБД – ситуация иная. Здесь все работы по комплектованию, систематизации, подготовке данных и др. иногда выполняются силами 2-3 человек и при этом не во всех МБД информацию индексируют специалисты-предметники. Квалификация персонала во многом зависит от уровня финансирования разработок систем и поэтому может быть разной в разных системах.

Предметная область документальной системы. Этот аспект рассмотрения проблем индексирования и разработки тезаурусов детально рассматривается в [3].

Далее на лингвистическом материале терминологически неразложимых КС в этой работе показывается, что критерии определения таких единиц, как и другие правила разработки тезаурусов, устанавливаются (или должны вырабатываться) с учетом взаимосвязи большого числа лингвистических, технологических, прагматических и других факторов или

условий работы систем разного типа и назначения.

Тезаурус по проблемам сохранности документов разрабатывался в соответствии с ГОСТ 7.25–80 [4]. Новые ГОСТы на эту работу в системе СИБИД, а именно ГОСТ 7.66–92, 7.70–96 и 7.74–96 при разработке тезауруса не использовались, так как были получены в БАН только в июле 2001г.

В ГОСТ на создание одноязычных тезаурусов имеется 6 критериев устойчивости, терминологической или поисковой неразложимости словосочетаний [4, с.3–4]. В учебном пособии ИПКИР [5, с.106–107] к ним добавлено еще несколько правил, определяющих основные особенности как «устойчивых», так и «свободных» лексических единиц ИПТ. Эти критерии и правила (ГОСТ и ИПКИР) отражают опыт работы большого числа крупных и малых документальных систем разного типа и назначения, имеют характер обычных для ГОСТ общих и мало конкретизированных рекомендаций, и по этой причине они не всегда дают индексаторам прямое и точное руководство в конкретных решениях проблемы определения устойчивых словосочетаний. Поскольку с середины 90-х гг. [6–7] такие рекомендации внедряются в практику индексирования информации в массовых библиотеках России, ниже дается их перечень и детальное описание того, как эти правила могут или не могут использоваться в работе специализированных, а также универсальных систем.

Согласно двум названным выше источникам, словосочетания не разбиваются, если они представляют собою такие лексические единицы (здесь и далее в скобках приводятся номера обсуждаемых критериев устойчивости). (1) Идиомы, то есть ЛЕ, содержание которых не сводится к сумме значений и не выводится из содержания их компонентов, например, *красная строка, царская водка* и т. п. (2) Термины с именем собственным; как *закон Ома, вольтова дуга* (3) Названия оборудования и материалов, пишущиеся через дефис (*смесители-запарники*).

Область использования трех выше названных критериев устойчивости языковых единиц, как правило, весьма ограничена. В БД ОСД, например, на эти три класса словосочетаний в настоящее время приходится только 10 наименований, а именно – *Hudson acidity с.Кислотноть по Хадсону, PAPER-FILM с.Система бумага-пленка, PAPER-GLUE с.Система бумага-клей* и некоторые др. Смысловая неразложимость первых двух классов ЛЕ (идиом и названий типа *распределение Пуассона, метод Дебая-Шерера*) едва ли у кого вызывает сомнение, и поэтому эти критерии устойчивости полезны лишь в незначительной степени в силу их очевидности. Третий критерий устойчивости – написание с дефисом

при сложно-сочинительном соотношении основ, по нашему мнению, имеет спорный характер и действует лишь при известных условиях работы определенных документальных баз данных, о чем говорится и в [5, с.107].

(4) Словосочетания типа *торговля на вынос, легкая промышленность,* определяемые как ЛЕ, «элементы [которых] не употребляются в составе других сочетаний или употребляются всегда в другом смысле» [4, с.3].С наших позиций, такие ЛЕ близко стоят к идиомам и едва ли их стоит рассматривать как какой-то особый класс устойчивых словосочетаний. Можно сказать, что наименования, описанные выше под пунктами 1–4, являются более или менее идиоматическими выражениями, это – «мелочи» ИПЯ, а их содержательные характеристики отражают лишь собственно лингвистический и далеко отстоящий от целей документального поиска аспект рассмотрения понятия устойчивости.

(5) Названия, имеющие в языке определенной области знания синонимы и аббревиатуры. Этот признак – один из важнейших в практике разработки тезаурусов, но он требует уточнения формулировки, поскольку речь может идти о разных синонимах и сокращениях. Очевидно, для каждого слова или устойчивого выражения ИПЯ можно найти или искусственным образом сконструировать синонимичное ему выражение. Многие аббревиатуры, например, в РЖ ВИНИТИ являются окказиональными, иногда создаются только для разового употребления в определенной статье или в ее реферате, а наличие у дескрипторов окказиональных синонимов и аббревиатур не может являться критерием их терминологической неразложимости. Имеются, с другой стороны, достаточно часто используемые сокращения вполне разложимых, свободных словосочетаний, как, например, LCSH и ПРБК – «предметные рубрики Библиотеки Конгресса». Таким образом, признаком неразложимых на составляющие их элементы ЛЕ является только общеизвестная синонимия и общепринятые сокращения, а также (6) наличие таких же ниже- и вышестоящих дескрипторов, когда «разбиение словосочетаний приводит к потере важных парадигматических связей, как АЛГОРИТМИЧЕСКИЕ ЯЗЫКИ – н.АЛГОЛ, КОБОЛ, ФОРТРАН» [4, с.4] и, следовательно, к информационным потерям при проведении поисков. Опора на парадигматику при индексировании не означает, что процесс индексирования информации во всех случаях должен идти с применением тезаурусов или каких-то иных списков нормализованных терминов. Речь в данном случае скорее идет о том, что в специальной литературе рассматривается под названиями «тезаурус пользователя», «тезаурус

индексатора» [8], то есть о некотором более или менее широком общекультурном комплексе знаний о самых разных вещах и взаимосвязях явлений.

(7) Словосочетания не расчленяются, если они называют измеряемые свойства, параметры, характеристики. Данное правило, скорее всего, имеет своим основанием потребности поиска и особенности индексирования информации в факто-документальных системах, как система, описанная в [9], где к названной категории слов в поисковых образах документов «привязываются» их числовые характеристики. В обычных БД, где у записей нет рефератов, содержащих фактографические данные, рассматриваемое правило индексирования информации является малополезным, а буквальное исполнение его, не ограниченное иными критериями, имело бы следствием то, что в ИПЯ базы данных пришлось бы ввести очень большое количество двух- и трехсловных словосочетаний. Так, для БД ОСД в настоящее время достаточно термина и глубины индексирования информации на уровне ПРОЧНОСТЬ (бумаги, картона и других материалов). Это понятие имеет большое количество нижестоящих, называющих измеряемые характеристики и образованных из таких составных элементов, как [ПРОЧНОСТЬ, ПРЕДЕЛ ПРОЧНОСТИ]-[НА СЖАТИЕ, ПРИ СЖАТИИ – НА РАЗРЫВ, ПРИ РАЗРЫВЕ – НА РАЗДАВЛИВАНИЕ, ПРИ РАЗДАВЛИВАНИИ и мн. др] – [ПРИ КОМНАТНОЙ ТЕМПЕРАТУРЕ, ПРИ ТЕМПЕРАТУРЕ 20 ГРАД. С, ПРИ ВЫСОКОЙ ТЕМПЕРАТУРЕ и др.]. Из выделенных скобками составных элементов ЛЕ, которые могут быть связаны с наименованием ПРОЧНОСТЬ, получается очень большое число синонимов и нижестоящих понятий только рассматриваемого нами дескриптора, а в системах, имеющих отношение к материаловедению, названия свойств иногда составляют значительный объем ИПЯ. Глубокое индексирование информации о свойствах, в связи с вышесказанным, оправдано главным образом лишь в условиях работы факто-документальных реферативных БД. В обычных библиографических базах данных (и особенно – в массовых библиотеках) рассматриваемое правило индексирования не дает никаких преимуществ при проведении поисков информации.

(8) Не делятся на составляющие их элементы ЛЕ, особенно часто используемые в какой-либо области знания (уксусная кислота, дозы облучения, товары широкого потребления). Этот критерий – несомненно существенный – имеет действительно универсальный характер и, весьма вероятно, имплицитно присутствует в большей части описываемых в

данной работе характеристик устойчивых языковых единиц. В теоретическом плане о связи понятия ‘устойчивость’ с частотностью терминов, интерпретируемой как вероятность совместного появления в текстах двух или более слов, писал И. А. Мельчук еще в 1960 г. [10]. В его понимании устойчивость является вероятностной характеристикой и принимает значения от 1 (чистые идиомы) до 0 (невозможные сочетания слов). Для решения проблемы устойчивости И. А. Мельчук предлагал массовую статистическую обработку большого количества текстов и расчет вероятностей (меры устойчивости) совместного употребления ЛЕ. Такое решение имеет, понятно, главным образом теоретический интерес, и практическая реализация поставленной в таком виде задачи является, видимо, невозможной из-за ее трудоемкости в исполнении даже для современной высокопроизводительной вычислительной техники, не говоря уже о методической стороне постановки и проведения данных расчетов. Но общий подход к обсуждаемому у И. А. Мельчука представляется вполне обоснованным. Специалисты-предметники и опытные индексаторы так или иначе оперируют вероятностными характеристиками, когда утверждают, что какое-то сочетание слов является редким или высокочастотным в документальном массиве системы. Частотный критерий устойчивости, как и другие критерии, должен использоваться не сам по себе, а во взаимосвязи с другими, рассматриваемыми в этом разделе. Так, на уровне компетенции обычного пользователя базы данных или носителя русского языка понятие ‘уксусная кислота’ ассоциируется, то есть находится в родовидовой подчинительной связи с понятиями ’кислоты’ и ‘химия’. Данная связь плюс высокая степень использования, и не только в документах по химии, и являются главным, чем, по нашему мнению, определяется терминологическая неразложимость словосочетания.

(9) Словосочетания, в которых имеются широкие по содержанию ЛЕ. Этот признак является слабым критерием устойчивости, и по нашей оценке, здесь дело не в том (или, точнее, не только в том), что в словосочетаниях типа «металлические конструкции», «математическое обеспечение» или «обеспечение сохранности документов» опорное слово имеет широкое содержание, а во вполне очевидных родовидовых отношениях, устанавливаемых в рядах таких терминов, как МЕТАЛЛИЧЕСКИЕ КОНСТРУКЦИИ н. [видовые понятия – конкретные виды металлических конструкций]; МАТЕМАТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ н. КОМПЬЮТЕРЫ, АВТОМАТИЗАЦИЯ, УПРАВЛЕНИЕ и др. Слова и словосочетания с общим (широким) значением, в силу такой

их семантики, в ИПЯ и в естественном языке называют понятия, относящиеся к «верхним» уровням иерархических связей определенных дескрипторов или просто лексических единиц. Поэтому анализируя данную лексику, иногда, даже не зная в деталях предметную область, можно с достаточной степенью обоснованности предполагать (прогнозировать) наличие у широких по содержанию ЛЕ подчиненных им нижестоящих понятий. Иерархия связей лексических единиц в данном случае устанавливается на основе того, что у Т. ван Дейка описано как сценарии, фреймы, общекультурный контекст [11], позволяющие рассматривать содержание понятий через призму определенных общеязыковых, когнитивных структур, причем независимо от специальных задач автоматизированного поиска информации.

Словосочетания разделяются, если в их содержании выделяются следующие элементы.

(10) Операция и объект операции, как «производство азота» – ПРОИЗВОДСТВО, АЗОТ [5]. У этого правила есть исключения и, в том числе, – наименования, относящиеся к общим уровням иерархии понятий в какой-либо области знания. В таких случаях объект операции часто обозначается не родительным падежом существительного, а относительными прилагательными, как, например, «огнеупорное производство», «сталелитейное производство» и др. Приведенные словосочетания называют не только (или не столько) определенный технологический процесс, сколько определенную отрасль или подотрасль в составе каких-либо более широких тематических областей, в данном случае – в составе предметной области «черная металлургия». Содержание данных понятий включает в себя не только аспект ПРОИЗВОДСТВО объектов, обозначаемых данными терминами, но также аспекты ИССЛЕДОВАНИЕ, ПРОЕКТИРОВАНИЕ и мн. др. Таким образом, можно сказать, что значение данных ЛЕ целиком не выводится из составляющих их элементов, и такие названия представляют собой близко стоящие к идиомам терминологически неделимые единицы.

Операция и объект операции иногда выражаются также грамматически нерасчлененно, одним сложным словом с подчинительным соотношением основ, как КНИГОТОРГОВЛЯ, ДОКУМЕНТОХРАНИЛИЩА или английские термины PAPERMAKERS, BOOKSELLERS, GOLDBEATERS и многие другие. Словосочетания PAPERMAKING [производство бумаги] и LEATHER MANUFACTURE [производство кожи], пока для них не были установлены «полезные» при проведении поисков нижестоящие, в соответствии с

описываемым правилом индексирования, разбивались на одиночные термины. В настоящее время такие дескрипторы употребляются в виде терминологически неделимых ЛЕ. Трудности с обработкой КС, как PAPERSTAINERS, PARCHMENT MAKERS и т. п. заключаются в том, что для некоторых таких слов не имеется эквивалентных им по содержанию русских однословных или двухсловных ЛЕ. По этой причине в тезаурусе по обеспечению сохранности документов есть такие не очень удачные номинации, как *Мастера по производству бумаги* см. PAPERMAKERS, *Мастера по производству чернил* см. INKMAKERS., *Торговцы акцидентным* [то есть сопутствующим] *книжным товаром* см. STATIONERS и т. п. Употребление данных КС имеет то неудобство, что рассматриваемые ключевые слова имеют характер искусственно образованных языковых конструктов, и поэтому хуже запоминаются индексаторами, чем обычные термины, их трудно отыскивать в словаре, и в синтаксически распространенных ЛЕ более вероятны ошибки в их написаниях (опечатки при записи информации).

Не расчленяются и такие словосочетания, как *Подготовка специалистов* см. EDUCATION; *Stains removal, Удаление пятен* см. CLEANING. Форма словосочетания для таких единиц была выбрана, среди прочего, потому, что синонимия ЛЕ в приведенных дескрипторах, вполне очевидная и легко устанавливаемая при ручном индексировании информации, при их расчленении не будет учитываться (контролироваться) при автоматическом индексировании документов. Таким образом, на решение вопроса об устойчивости или свободном характере словосочетаний в тезаурусе влияют также особенности программного обеспечения системы.

(11) Часть и целое, например, «катоды радиоламп» – КАТОДЫ, РАДИОЛАМПЫ. – Исключений из этого правила имеется небольшое количество. В БД ОСД к ним относятся, например, дескрипторы, называющие части книги, как *Корешки книжных блоков* см.BACKS, *Поля документов* см.MARGINS и нек. др. В таких случаях русские словосочетания имеют синонимичные им однословные английские термины.

Объект и его назначение: «станки для распилки карандашей»: – СТАНКИ, РАСПИЛКА, КАРАНДАШИ. – Такие наименования могут иметь синонимы, как, например, «элемент, реагирующий на уровень жидкости» – МЕМБРАНЫ, ПОПЛАВКИ [12], «устройства для перекрытия струи металла» – СКОЛЬЗЯЩИЕ ЗАТВОРЫ. Поэтому прежде чем разбивать описательные выражения на составляющие их элементы, необходимо

убедиться вначале, что у данных ЛЕ не имеется терминологизованных обозначений и, если такие наименования есть, то провести операцию терминирования, то есть замену описательных выражений на термины. По нашему мнению, вполне допустимы, не рекомендуемые в [13], синонимы типа УСТРОЙСТВА ДЛЯ ШТАБЕЛИРОВАНИЯ при дескрипторе ШТАБЕЛЕРЫ. Такие лексические единицы могут употребляться тогда, когда индексаторам, например, не известно наименование ШТАБЕЛЕРЫ, или же это название не сразу «приходит на память».

По формальным, а не содержательным признакам также рекомендуется разделять:

(13) сочетания двух существительных, как «история России» – ИСТОРИЯ, РОССИЯ – без какой-либо характеристики этого случая в рассматриваемом нами пособии [5];

(14) сочетания существительных и прилагательных, если они не являются устойчивыми терминами в политематической области: «военная доктрина» – ВОЕННЫЙ, ДОКТРИНА; «опорная конструкция» – ОПОРА, КОНСТРУКЦИЯ. Содержание данного признака терминологически неделимых ЛЕ является также не очень понятным, как и в выше указанном случае;

(15) словосочетания, содержащие два или более прилагательных. Такие словосочетания заменяются на несколько двухсловных ЛЕ, в которых одно и то же существительное сопровождается поочередно одним из прилагательных, как, например, «торированные вольфрамовые катоды» – ТОРИРОВАННЫЕ КАТОДЫ, ВОЛЬФРАМОВЫЕ КАТОДЫ [7, с.44].

И последний критерий определения терминологически неразложимых ЛЕ, сформулированный в [5] таким образом:

(16) «Решения о разделении или сохранении ЛЕ словосочетания целесообразно принимать, исходя из интересов поиска информации для каждой составной ее части» [с. 107]. Очевидно, что речь здесь идет об учете при составлении тезаурусов информационных потребностей пользователей.

Использовать последний критерий на практике затруднительно в силу того, что методы определения информационных потребностей (например, анкетированием) не совершенны, а успешность решения данной задачи во многом зависит от класса и общего назначения документальной системы. В специализированных МБД основные потенциальные информационные потребности пользователей устанавливаются более или менее легко и основная их часть определяется при рассмотрении документов и лингвистического материала системы, иногда с помощью

небольших, но хорошо покрывающих предметную область массивов профильной для БД информации. В универсальных документальных системах задача определения информационных потребностей, как и устойчивости языковых единиц, является более сложной.

Наконец приведем наши собственные критерии определения устойчивых словосочетаний (и не только устойчивости единиц ИПТ). Эти критерии могут быть сформулированы в виде двух принципов – принципа единообразия, параллелизма и аналогии в принимаемых лингвистических решениях и правила экономии лингвистических средств или минимизации тезауруса.

(17) Под единообразием в данном случае мы имеем в виду не единообразное индексирование информации как важнейший из показателей качества выполнения индексационных работ, но одинаковые или близкие типовые решения при обработке определенных видов ЛЕ. Критерий единообразия основывается на использовании при составлении ИПТ небольшого числа обобщенных лексико-семантических категорий, какими в БД ОСД являются категории ‘операция’, ‘процесс’, ‘вещество’, ‘оборудование’ и некоторые др. Примерами дескрипторизации лексики, основанными на рассматриваемом принципе, могут являться такие дескрипторы, как FREEZE DRYING [сушка вымораживанием] н.BLAST FREEZERS [устройства для сушки обдувом], LEAFCASTING [доливка] н.LEAFCASTING MACHINES [доливочные машины], DISINFECTION н.DISINFECTION CHAMBERS. Общим у данных дескрипторов является общий концепт ‘операция’, а различия заключаются в том, что в одном случае данный концепт обозначается просто как операция, а в другом – как принцип работы устройства или его назначение. В тезаурусе между такими дескрипторами почти во всех случаях устанавливается “типовая” родовидовая парадигматическая связь, при которой слова с категориальным значением ‘оборудование’ являются нижестоящими по отношению к ЛЕ, называющим соответствующие операции. Такая же регулярная связь устанавливается, когда родовое понятие относится к категориям ‘операция’ или ‘процесс’, а видовое представлено наименованием вещества, каким-либо образом с ними связанного. Например: BLEACHING [отбеливание] н.BLEACHING AGENTS [отбеливающие вещества], DISINFECTION н.DISINFECTION AGENTS, NEUTRALIZATION н.NEUTRALIZATION AGENTS и т.п.

С использованием принципа аналогии в БД ОСД образованы также дескрипторы, как,

NATIONAL LIBRARIES OF FINLAND

с. Finnish national libraries
Национальные библиотеки Финляндии
Финские национальные библиотеки

в. NATIONAL LIBRARIES

н. HELSINKI LIBRARY
NATIONAL LIBRARY OF HEALTH SCIENCE IN FIN LAND

В данном случае аналогия прослеживается не в категориальной семантике синонимичных ЛЕ, а в грамматических формах, в которых представлены наименования документохранилища. В дескрипторах данного типа должны содержаться:

- английский дескриптор, в котором на первом месте находится прилагательное, называющие вид или общее назначение документохранилища. Например: NATIONAL, UNIVERSITY, PUBLIC, SPECIAL и т. д. [LIBRARIES, ARCHIVES, MUSEUMS и др.];
- такой же русский синоним, как, например, «Национальные библиотеки» [музеи, архивы, ассоциации и др.] Великобритании [Франции и т. д.];
- английский синоним, который начинается с прилагательного, называющего географическую привязку объекта, в данном случае – документохранилища. Например: «British library», «Russian museum», «Australian national library», «New York public library» и т. п.;
- русский синоним, построенный по такой же семантико-синтаксической модели.

Приведенная унификация единиц ИПТ на основе критерия аналогии и сходства существенно упрощает работу с тезаурусом и делает более контролируемым процесс индексирования информации. При этом категориальный (фасетный) анализ является, разумеется, не единственным при составлении ИПТ. Широко применяются также обычные классификации понятий в предметной области ОСД. Последние, как и любые другие научные классификации, в свою очередь, также основываются на обобщенных понятиях (категориях) типа «сырье», «вещество», «материал», «оборудование», «процесс», «операция» и т. п.

(18) Экономия лингвистических средств базы данных имеет место тогда, когда, принимая какие-то правила, разработчики ИПТ избегают решений, ведущих к малооправданному и при этом существенному увеличению объема тезауруса. На принципе минимизации тезауруса, по нашему мнению, основываются и некоторые из выше указанных критериев

устойчивости языковых единиц. Так, описанное в пункте 13 разделение словосочетания «история России» на составляющие его элементы скорее всего обусловлено тем, что понятия ‘история’ и ’Россия’ как самостоятельные элементы ИПЯ могут быть также использованы для индексирования очень широкого круга запросов, как, например, «история Англии, Франции, Португалии и т. д.» «экономика, население, финансы, военное дело и мн. др. в *России* 18 века» В дескрипторных ИПЯ связь понятий, называемых одиночными терминами, осуществляется не заранее, до проведения поисков информации, как в сложных предметных рубриках каталогов, а при помощи координации терминов в поисковых выражениях запросов. По этой причине в дескрипторных ИПЯ ключевые слова, содержащие три и более элемента, являются часто ненужными, так как имеются более «экономные» языковые средства индексирования информационных запросов. Содержание многих запросов описывает семантико-синтаксическая модель *[предмет индексирования информации] – [аспект] – [характеристика места] – [характеристика времени]*, и эта модель широко применяется для составления рубрик предметизационных ИПЯ. В связи с обсуждаемым нужно специально сказать, что дескрипторные и предметизационные языки индексирования – это разные классы ИПЯ. Соотношение между ключевыми словами и сложными предметным рубриками, по образному определению Ф. И. Грининой, такое же, как соотношение между кирпичами (ключевые слова) и кирпичными строительными конструкциями (сложные предметные рубрики) [14], а смешение двух разных принципов разработки тезаурусов неправомерно как теоретически, так и в практическом плане. И поэтому к выше описанным критериям устойчивости языковых единиц можно добавить еще один – принципиальный выбор типа ИПЯ базы данных. Если при разработке документальной системы был выбран предметизационный ИПЯ, то соображения о малооправданной сложности языка индексирования информации теряют свою актуальность и становятся малосущественными.

Экономия лингвистических средств как критерий устойчивости «действует», как и все остальные, описанные выше, только с учетом взаимодействия многих условий и факторов, и не является главным и определяющим. Так, трехсловное сочетание УЧЕТ БИБЛИОТЕЧНЫХ ФОНДОВ в отношении его синтаксической формы (но не содержания и употребления) – такое же, как и «свободное» словосочетание «история России». Однако в отличие от последнего эта ЛЕ: (1) относительно часто

встречается в литературе по обеспечению сохранности документов, (2) является общепринятым термином, (3) имеет в БД ОСД вышестоящий дескриптор PRESERVATION [обеспечение сохранности документов], и поскольку близких к нему в семантико-синтаксическом отношении ЛЕ, кроме синонима *комплектование фондов*, в тезаурусе по ОСД не имеется, по принципу аналогии (4) не может быть образовано большое число таких же, но малополезных при проведении поисков терминов. Приведенное комплексное рассмотрение «бытования» термина *учет библиотечных фондов* и позволяет определить его как терминологически неделимую единицу.

Из рассмотрения проблемы устойчивости, как одной из центральных проблем, решаемых при разработке тезаурусов, вытекает, что в ходе дескрипторизации лексики разработчики ИПТ руководствуются не только небольшим числом правил (в ГОСТ на создание одноязычных тезаурусов их описано, например, только 6) установления терминологически неразложимых ЛЕ, простых и понятных любому носителю русского языка с известным уровнем общего образования, но и учитывают много критериев, часть которых далеко отстоит от собственно лингвистической стороны рассмотрения проблемы. Правила индексирования информации и составления ИПТ представляют собою не «жесткие» и не знающие исключений закономерности, как это имеет место в точных науках или в химии, физике и других дисциплинах естественнонаучного профиля. Исключения из рассматриваемых правил обусловлены тем, что поскольку данные правила вырабатываются с учетом большого количества факторов в конкретных условиях разработки или эксплуатации информационных систем, не все они могут учитываться в полном объеме, и какие-то из критериев объективно являются (или оцениваются) как более важные и «весомые», чем другие. Отсюда – и неизбежная субъективность и неоднозначность в конкретных решениях вопроса о терминологической разложимости/неразложимости языковых единиц в различных документальных системах.

Литература

1. Соловушкова Г. Э., Пименов Е. Н., Амхир И. К. Опыт работы автоматизированной информационно-поисковой системы по огнеупорам// Огнеупоры. – 1989. – № 3. – С. 43 – 48.
2. Horny K. Cataloging simplification: trends and prospects// Int. cat. and bibliogr. control. – 1991. – Vol. 20, № 3. – С. 25 – 28.

3. Пименов Е. Н. О факторах, влияющих на индексирование: индексирование и предметная область // НТИ. Сер.1. – 2000. – № 2. – С. 15 – 23.
4. ГОСТ 7.25–80. Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления. Введен. 01.01.82 – Изм.1. Вед. 01.07.87.
5. Воройский Ф. С. Аналитико-синтетическая обработка и переработка информации в автоматизированных системах НТИ: Основы организации и технологии. (Учебное пособие). – М.: ИПКИР, 1991. – 289 с.
6. Воройский Ф. С. Некоторые пути повышения качества поисковых характеристик электронных каталогов// Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества: Материалы конф., 1-9 июня 1996 г. – М., 1996. – С. 286 – 289.
7. Воройский Ф. Индексирование документов в АБИС // Библиотека. – 1996. – № 9. – С. 42 – 44.
8. Садовска Я. Проблемы предметного поиска в on-line каталогах // Предметный поиск в традиционных и нетрадиционных информационно-поисковых системах. – Вып.11 – СПб: РНБ, 1994. – С.108.
9. Методика индексирования запросов в АСНТИ-СМ: Методические рекомендации. – М.: ВИИЭСМ, 1981. – 15 с.
10. Мельчук И. А. О терминах «устойчивость» и «идиоматичность»// Вопросы языкознания. – 1960. – № 4. – С. 73 – 80.
11. ван Дейк Т А. Язык. Познание. Коммуникация: Сб. работ. – М.: Прогресс, 1989. – 312 с.
12. Методика индексирования документов по «Тезаурусу по атомной науке и технике» для системы автоматизированного распределения информации. – М: ЦНИИатоминформ, 1977. – 49 с.
13. Ханжин А. Г. Разработка методики координатного индексирования информации. Часть 2 // НТИ. Сер.2. – 1995. – № 9. – С. 14 – 17.
14. Гринина Р Ф. К вопросу о соотношении дескрипторных и предметизационных языков // Труды Гос. ин-та культуры им. Н. К. Крупской. – XXIII – Л., 1970. – С. 53 – 63.